PROBABILISTIC GENERATIVE MODELING OF SPEECH

BY

YANG ZHANG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Adviser:

Professor Mark Hasegawa-Johnson

ABSTRACT

Speech processing refers to a set of tasks that involve speech analysis and synthesis. Most speech processing algorithms model a subset of speech parameters of interest and blur the rest using signal processing techniques and feature extraction. However, evidence shows that many speech parameters can be more accurately estimated if they are modeled jointly; speech synthesis also benefits from joint modeling.

This thesis proposes a probabilistic generative model for speech called the Probabilistic Acoustic Tube (PAT). The highlights of the model are threefold. First, it is among the very first works to build a complete probabilistic model for speech. Second, it has a well-designed model for the phase spectrum of speech, which has been hard to model and often neglected. Third, it models the AM-FM effects in speech, which are perceptually significant but often ignored in frame-based speech processing algorithms. Experiment shows that the proposed model has good potential for a number of speech processing tasks.

To my parents, for their love and support.

ACKNOWLEDGMENTS

I would like to acknowledge my graduate advisor, Professor Mark Hasegawa-Johnson, who has given me lots of research opportunities, guidance and insights. His broad knowledge and research attitude deeply cultivated me to become an independent, innovative and upright researcher.

I would also like to acknowledge my undergraduate advisor, Professor Zhijian Ou, who initiated the idea of the Probabilistic Acoustic Tube model and worked on it continuously. His innovative ideas often pushed the project forward.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Motivation of Joint Modeling

Speech analysis/synthesis refers to a family of speech processing applications, such as speech modification, coding, enhancement, and recognition [1]. Most speech analysis/synthesis systems are based on the basic physical model of speech production - the acoustic tube model, also known as the source-filter model [1].

To better study speech processing techniques and speech modeling, it is useful to take a look at how speech is produced. Figure 1.1 shows an anatomic view of the human speech system [1]. The lungs push air through the trachea, and the air passes the vocal folds, which modulate it into a quasi-periodic signal, normally called a glottal wave, or vocal excitation. The vocal excitation then passes through the vocal tract, which consists of oral cavity and nasal cavity. The articulators of the vocal tract, such as tongue, jaw and teeth, are placed in certain positions to form some resonance frequencies. These frequencies are called formant frequencies, which are very important for speech recognition. The filtered sound wave is emitted at the lips and radiates, becoming what we call speech.

The paragraph above describes speech excited by vocal fold vibration (voiced speech), which dominates, both in energy and duration, human utterances and mostly corresponds to vowels. In other cases, however, air flow does not get modulated by the vocal folds before it passes through the vocal tract, and forms unvoiced speech, which roughly corresponds to consonants. In some situations, e.g. plosives and fricatives in English, some part of the vocal tract gets contracted, forcing high speed, irregular turbulent air flow.

If we define the vocal tract as a system, and glottal vibration as the source, then the production system just described constitutes a source-filter model.

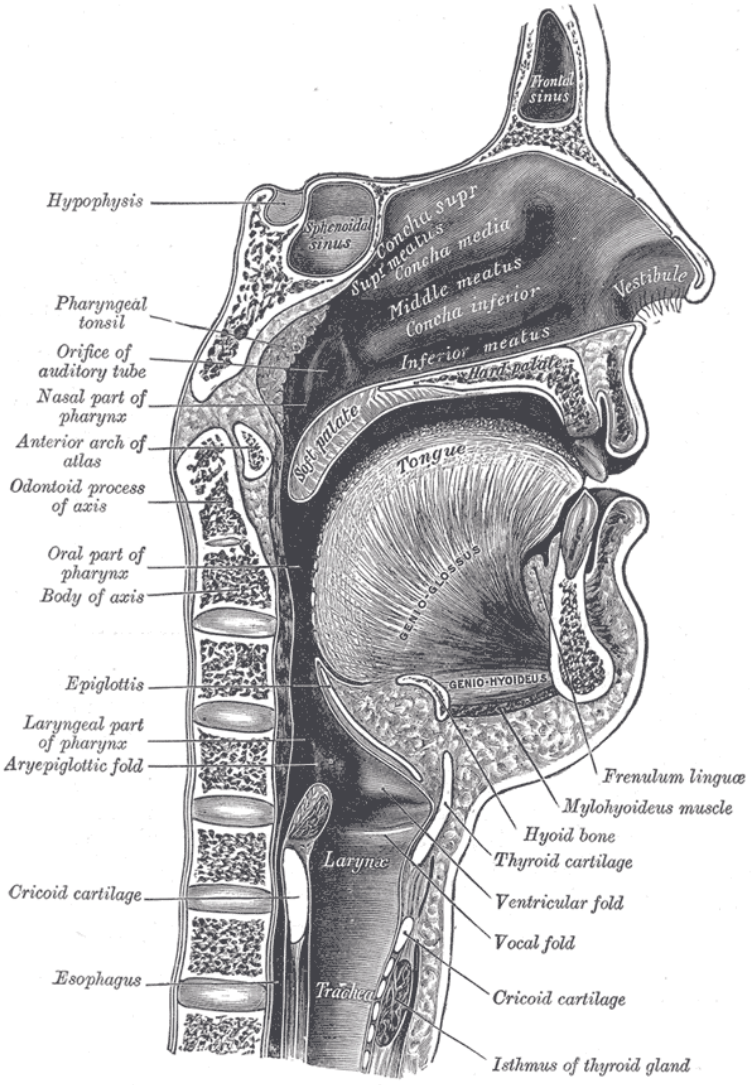


Figure 1.1: Human speech system[a]

[a] "Sagittalmouth". Licensed under Public Domain via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Sagittalmouth.png#/media/File:Sagittalmouth.png

Therefore, speech, as the output of the source-filter model, contains the information of the source, including pitch, GCI, glottal wave and aspiration, and the filter, which is the vocal tract.

In analysis, most tasks only focus on a part of the above information. Therefore, a common design paradigm is to build a special-purpose signal-processing front-end that extracts the most relevant features for the target task while suppressing the interference induced by the rest. Here we give two examples.

Pitch Tracking

Pitch tracking essentially extracts the voiced/unvoiced states. For voiced speech, it further estimates F0, which is the frequency of the glottal wave.

A common approach to pitch tracking [2, 3, 4, 5, 6, 7] is through autocorrelation function $R(\tau)$, defined as

$$R(\tau) = \sum_t s(t)s(t-\tau) \tag{1.1}$$

where $s(t)$ is the (framed) time-domain speech signal. It can be easily shown that $R(\tau)$ achieves a maximum at $\tau = 0$, but for a voiced signal with fundamental period T_0, $R(\tau)$ will achieve local maxima at $\tau = kT_0$ for integer k.

A problem of this approach is false peaks, i.e. besides peaks at multiples of T_0, there are many other peaks, which can be easily mistaken as the peaks at multiple pitch periods. To alleviate this problem, some research studies perform center clipping of the speech signal before calculating its autocorrelation [8].

But the key to this problem lies in the interference of the glottal wave shape and the filter. Center clipping itself is essentially a method to remove the interference.

Speech Recognition

Vowel identity and consonant place of articulation are encoded largely in the filter, or the vocal tract transfer function, which roughly shows in the spectrum as the spectral envelope. A common method to estimate spectral envelope is MFCC [9], which essentially blurs the fine structure by smoothing, which suppresses the pitch information.

However, the drawback of this paradigm is that the omitted information, though suppressed, can still bring significant interference, whereas in fact it

can actually help the target task if properly modeled. For example, Kameoka et al. [10] noted that pitch and spectral envelope have a chicken and egg relationship and should be estimated jointly. Stephenson [11] pointed out that cepstral-based features are sensitive to auxiliary information such as pitch and energy.

Previous work also showed evidence of glottal wave impacting on spectral envelope. For example, Klatt and Klatt [12] point out that different glottal activities (breathy, modal, laryngealized) would introduce different levels of spectral tilt and positions of glottal formant. Therefore vocal tract and glottal information can be more accurately estimated if considered together. Therefore, speech processing tasks can be greatly benefited if a complete model of speech that jointly models all the aforementioned speech parameters is available.

1.2 Previous Work on Joint Modeling

There are a few existing works on joint modeling of speech parameters. Some works focuses on deconvolution of the source excitation and filter for high-quality speech reconstruction and manipulation. For example, Degottex et al. [13] proposed a speech model called SVLN, using pitch, glottal source and vocal tract as its main variables. The STRAIGHT model [14, 15] is a source-filter-based speech model for speech modification. It jointly models pitch, spectral envelope and aspiration. Achan et al. [16] proposed a time-domain probabilistic speech model that infers the excitation and the impulse response jointly from speech. Kameoka et al. [17] proposed a harmonic temporal structured clustering (HTC) method that jointly models the harmonic structure (excitation) and spectral envelop, which can be used for speech reconstruction and other tasks.

Other related works focus on fine modeling of glottal activities such as glottal wave and aspiration. For example, Jackson and Shadle [18] proposed a joint model for voiced and unvoiced excitation. Alku [19] proposed an iterative algorithm that simultaneously estimates the glottal wave and vocal tract response. Drugman et al. [20] proposed a causal-anticausal decomposition scheme that jointly estimates the vocal tract and glottal wave.

In speech enhancement, increasing attention has been paid to apply joint

models of speech to improve naturalness of the enhanced signal. The NMF-based [21, 22, 23, 24, 25] and ICA based [26, 27, 28, 29, 30] source separation blindly decompose speech into a base matrix, which can be interpreted as the excitation, and a coefficient matrix, which can be interpreted as the system. Tirumala and Mandel [31] and Mandel et al. [32] propose a source-filter-based speech denoising algorithm that obtains noise-robust estimates of pitch and spectral envelope and resynthesizes clean speech using the estimates.

In speech synthesis, jointly modeling glottal wave, aperiodicity and vocal tract is becoming an increasingly popular approach to improve naturalness of synthetic speech. For example, Rosenberg [33] noted that using different glottal waves for synthesis results in differences in perception and subjective preference. Raitio et al. [34, 35] proposed a HMM-based parametric speech synthesizer that builds a library of glottal waves obtained from a speech vocoder. Maia et al. [36] and Sang-Jin and Minsoo [37] proposed a mixed-excitation synthesis system that jointly models voiced and unvoiced excitation to improve the naturalness of synthetic speech. Cabral et al. [38] proposed a HMM-based speech synthesis system that uses the LF-model [39] for the glottal source.

There are three major limitations regarding the existing works, partially due to the scope of their target applications. First, some of these models still mix some speech parameters. For example, the STRAIGHT model and Tirumala's denoising scheme still mix glottal wave and vocal tract response.

Second, although the models jointly consider different speech parameters, the estimations of these parameters are still performed separately, which, as previously discussed, still suffers from mutual interference.

Finally, these models either neglect or only partially model phase. In particular, in most NMF- and ICA-based source separation approaches, only the magnitude spectrogram is modeled. The estimated separated signal is obtained by masking [40] on the magnitude spectrum and directly applying the phase spectrum of the mixture signal. This paradigm is the most common scheme for other separation algorithms, including deep learning [41, 42] and probabilistic models [43, 44]. This paradigm suffers from residual noise, also called music noise [45], and artifacts.

1.3 The Probabilistic Acoustic Tube (PAT) Model

In this thesis, we propose a probabilistic generative model of speech, called the Probabilistic Acoustic Tube (PAT) model [46, 47, 48]. There are several highlights regarding this model.

First, it is among the very first works to build a complete probabilistic model for speech. In particular, it jointly considers pitch, glottal wave, glottal closure instance (GCI), aspiration and vocal tract response as hidden variables. The model can potentially be applied to both speech analysis and synthesis. For speech analysis, statistical inference techniques are applied to jointly infer the hidden variables, which makes the model unlike the separate estimation in existing works. For speech synthesis, the values of the hidden variables are specified as input to the generative model, whose output is thus synthetic speech. Our study in this thesis is in spirit similar to the generative modeling approach to computer vision [49] that successfully accounts for different sources of variability in images and relies on learning and inference to perform various image analysis tasks. We demonstrate the capability of PAT for a number of speech analysis/synthesis tasks, such as pitch tracking under both clean and additive noise conditions, speech synthesis, and phoneme clustering.

Second, it has a well-designed model for the phase spectrum of speech, which has been hard to model and often neglected. The difficulty of phase modeling lies in aliasing and its poor noise robustness. One of the traditional approaches to phase modeling is phase unwrapping [50], but this method fails when SNR is low. The PAT model overcomes the difficulty by properly parameterizing the complex spectrum of each speech component, which has been well-studied over the past century. This idea is in principle similar to a number of speech models, such as the mixed-phase model [51, 52], but we incorporate this idea in a probabilistic generative manner.

Third, it models the AM-FM effects in speech, which are perceptually significant [1] but ignored in frame-based speech processing algorithms. Traditional approaches to AM-FM modeling/analysis of speech include Hilbert transform [53], sinusoid models [54] and probabilistic amplitude and frequency demodulation (PAFD) [55, 56]. However, it is hard to incorporate these approaches into the frame-based probabilistic framework of PAT. The PAT model approximates the stochastic AM-FM behavior with multivariate nor-

mal distribution by similar assumptions to those in Bayesian spectral estimation (BSE) [57].

1.4 Thesis Organization

The remainder of the thesis is organized as follows: chapter 2 introduces the relevant signal processing theories on speech and formulates the signal model of PAT; chapter 3 describes the probabilistic model of PAT by introducing probabilistic assumptions on the signal model introduced in chapter 4; and chapter 5 concludes the thesis and discusses future research directions.

CHAPTER 2

SIGNAL MODELING OF PAT

The signal model of PAT is based on some classical speech signal processing theories. This chapter goes through these theories before introducing the signal model of PAT.

2.1 Notations

Before the signal model is formally introduced, it is useful to define the notations that will be frequently used within this section.

Lower case letters with parentheses, such as $h(t)$, denote discrete time domain signals. Upper case letters with parentheses, such as $H(\omega)$ and $H(z)$, denote the DTFT and Z-transform respectively. Lower case letters with brackets, such as $h[n]$, denote cepstrum. $\mathcal{Z}^{-1}(\cdot)$ denotes inverse Z-transform operation; DTFT$(\cdot)$ denotes DTFT operation. $\circledast$ denotes circular convolution.

2.2 The Source-Filter Model

Speech can be modeled as the output of a source-filter model, where the source is glottal vibration and aspiration, and the filter is the vocal tract.

To show that the vocal tract can be modeled as a filter, we need to show it is linear and time-invariant (LTI) within a short time period.

First, within a short time frame, typically 30 ms, articulators move little, and therefore the system response can be regarded as time invariant. Second, if we assume that the air velocity $v(x,t)$ inside the vocal tract is small, the viscosity is negligible, the air density ρ remains constant, and the air only moves along the axial direction of the vocal tract, then the air pressure $p(x,t)$

and velocity satisfy the following set of linear equations [58, 59]:

$$\begin{aligned} -\frac{\partial p}{\partial x} &= \rho \frac{\partial v}{\partial t} \\ -\frac{\partial p}{\partial t} &= \rho c^2 \frac{\partial v}{\partial x} \end{aligned} \tag{2.1}$$

which is linear w.r.t. $v(x,t)$ and $p(x,t)$.

The LTI nature enables us to apply linear system theories to process speech [60]:

$$S(z) = E(z)H(z) \tag{2.2}$$

where $S(z)$, $E(z)$ and $H(z)$ are Z-transforms of the speech signal, the excitation and the system respectively.

The study of speech boils down to the study of the system, i.e. the vocal tract, and the excitation, i.e. the glottal wave.

2.3 Vocal Tract and Radiation

The oral tract, which dominates the vocal tract, can be modeled by a concatenation of P hard, lossless uniform tubes with different area A_k. It can be shown that if we sample the signal by $\tau = 2L/Nc$, where L is the total length of the oral tract, the vocal tract system is approximately all-pole [60], namely

$$H(z) = \frac{G}{1 - \sum_{k=1}^{P} \alpha_k z^{-k}} \tag{2.3}$$

and that the radiation at the lips can be approximated by 1st-order difference [61], namely

$$R(z) = 1 - z^{-1} \tag{2.4}$$

where $R(z)$ is the transfer function of radiation.

Equation (2.3) implies that the vocal tract system can be well modeled by an all-pole system. It can be shown that if the reflection coefficient between the k-th and (k+1)-th tube

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \tag{2.5}$$

is less than 1, which holds in reality, then the system is stable, i.e. all the

poles are within the unit circle.

If we take into account the impact of the nasal tract, which can be modeled as an all-pole system similarly, then the transfer function becomes

$$\begin{aligned} H(z) = H_o(z) + H_n(z) &= \frac{G_o}{P_o(z)} + \frac{G_n}{P_n(z)} \\ &= \frac{G_n P_o(z) + G_o P_n(z)}{P_o(z) P_n(z)} \end{aligned} \tag{2.6}$$

where $H_o(z)$ and $H_n(z)$ are transfer functions of oral tract and nasal tract respectively, $P_o(z)$ and $P_n(z)$ are their pole polynomials. The system is no longer all-pole. The poles still fall inside the unit circle. It can be proved that the zeros, which are the roots of $G_n P_v(z) + G_v P_n(z)$ also fall inside the unit circle. Therefore the resulting system is minimum-phase.

2.4 Cepstral Analysis

The minimum-phase $H(z)$ can be well-parameterized by cepstral coefficients [62], which is the most popular feature for speech and speaker recognition systems [63, 64, 65, 66]. While a more detailed derivation is given in [67], here we give a brief overview of the theory. Rewrite equation (2.6) into the pole-zero representation

$$H(z) = \frac{\prod_{k=1}^{N} (1 - n_k z^{-1})}{\prod_{k=1}^{P} (1 - p_k z^{-1})} \tag{2.7}$$

where N and P are number of zeros and poles respectively; n_k and p_k are the k-th zero and pole respectively. Since the system is minimum-phase, $|n_k|$ and $|p_k|$ are both smaller than 1.

Taking the logarithm of both sides and applying Taylor expansion, we have

$$\begin{aligned} \log H(z) &= \sum_{k=1}^{N} \log\left(1 - n_k z^{-1}\right) - \sum_{k=1}^{P} \log\left(1 - p_k z^{-1}\right) \\ &= \sum_{n=0}^{\infty} \left[\sum_{k=1}^{N} \frac{(-n_k)^n}{n} z^{-n} - \sum_{k=1}^{P} \frac{(-p_k)^n}{n} z^{-n} \right] \end{aligned} \tag{2.8}$$

The complex cepstrum is defined as the inverse Z-transform of $\log H(z)$.

According to the definition of Z-transform, we have

$$\begin{aligned} h[n] &= \mathcal{Z}^{-1}\left(\log H(z)\right) \\ &= \begin{cases} \frac{\sum_{k=1}^{N}(-n_k)^n - \sum_{k=1}^{P}(-p_k)^n}{n} & \text{if } n \geq 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{2.9}$$

As can be seen, $h[n]$ decreases at the rate of at least $1/n$, and is right-sided. Therefore, the vocal tract system can be parameterized by the cepstral coefficients at low quefrency, i.e.

$$H(\omega) = \exp\left[\text{DTFT}(\hat{h}[n])\right] \tag{2.10}$$

where $\hat{h}[n]$ is the truncated and zero-padded $h[n]$:

$$\hat{h}[n] = \begin{cases} h[n] & \text{if } n \in [0, C] \text{ for some } C > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

2.5 Glottal Wave

In the voiced case, the glottal wave excites the vocal tract system. The glottal pressure wave is a quasi-periodic signal, whose circular frequency $\omega_0 = 2\pi F_0/F_s$ is called the fundamental frequency. Figure 2.1 top panel shows a typical glottal pressure wave $g[n]$ within each period. As can be seen, there are three phases: glottal open phase, glottal return phase, and glottal closed phase. The instant of maximum derivatives between the open phase and return phase is called GCI (glottal closure instant). The periodic signal can be expressed as a periodic pulse train $p[\tau]$ convolved with some impulse response $g[\tau]$, in which $g[\tau]$ is compactly represented if one places impulses at the locations of the GCI, and thus the time of the first GCI is considered as the group delay.

According to the sampling-periodic duality, the DTFT of the periodic glottal wave $S_g(\omega)$ can be denoted as

$$S_g(\omega) = aP(\omega)G(\omega) \tag{2.12}$$

where $P(\omega)$ is a pulse train with interval $1/\omega_0$, $G(\omega)$ roughly corresponds to

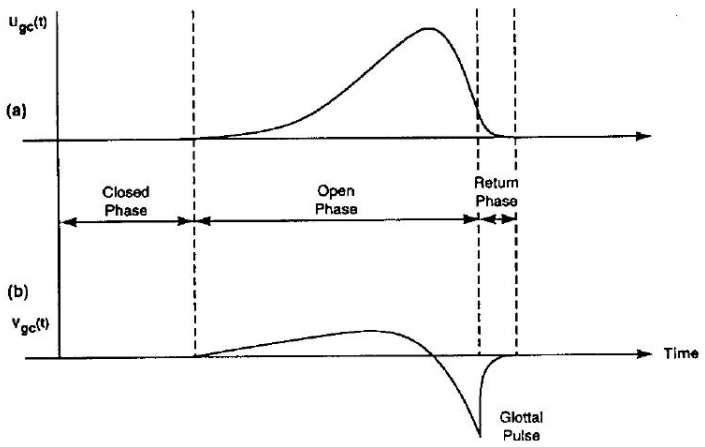


Figure 2.1: Glottal flow (upper panel) and its derivative (lower panel)[68].

the DTFT of the glottal wave within a single period, and a is the amplitude. Further, denote the DTFT of the output speech wave as $S_V(\omega)$. According to the convolution theorem, we have the following relation:

$$S_V(\omega) = aP(\omega)G(\omega)H(\omega)R(\omega) \tag{2.13}$$

By convention, the radiation effect and glottal wave are merged; namely, we often regard $E_V(\omega) = P(\omega)G(\omega)R(\omega)$ as the equivalent voiced excitation of the system $H(\omega)$. Then we have

$$S_V(\omega) = aE_V(\omega)H(\omega) \tag{2.14}$$

In the time-domain, this can be written as

$$s_V(t) = \sum_d |\alpha_d| \exp\left(j\left(d\omega_0(t-\tau) + \angle\alpha_d\right)\right) \tag{2.15}$$

where d is harmonic number, τ is the group delay of $P(\omega)$, i.e. the time instance when the first pulse occurs, $|\cdot|$ and $\angle$ denote magnitude and angle of a complex number, respectively, and

$$\alpha_d = aG(d\omega_0)H(d\omega_0)R(d\omega_0) \tag{2.16}$$

Notice that $R(\omega)$ is a first-order difference operator, so $E_V(\omega)$ is essentially the 1st-order difference of the glottal wave $G(\omega)$. A typical $E_V(\omega)$ in a single

period is shown in figure 2.1 bottom panel.

There have been research efforts to derive a parametric glottal model. For example, Holmes [69] and Michaels et al. [70, 71] studied the waveform shapes obtained by inverse filtering and high-speed motion pictures.

The LF model [39] is so far the most popular model of glottal wave within 1 period. It can be shown that the LF model can be further simplified as a three-pole model [72, 52]. According to the three-pole model, $E_V(\omega)$ can be characterized as passing a periodic pulse train $P(\omega)$ to an ARMA system $G(\omega)R(\omega)$ with a pair of anti-causal poles, which correspond to the opening phase, and a causal pole, which corresponds to the return phase. Formally,

$$G(\omega)R(\omega) = \frac{1}{(1 - 2g_1\cos(\beta_1)\exp(-j\omega) + g_1^2\exp(-2j\omega))\,(1 - g_2\exp(-j\omega))} \tag{2.17}$$

where g_1 and β_1 are the norm and absolute value of angle of the anti-causal pole pair, and g_2 is the real part of the causal pole.

The frequency representation of the glottal system, $G(\omega)R(\omega)$, is characterized by a distinct energy band in low frequency, called the glottal formant, and decreasing energy as frequency increases, called spectral tilt. Therefore, $S_V(\omega)$ is low-passy and contains a glottal formant.

It should be noted that these glottal wave models only capture the coarse structure of the glottal wave, but in actuality there exist ripples and fluctuations in the glottal wave, which introduce high frequency variations [68, 73].

In unvoiced cases, the air wave does not vibrate the vocal chords. A typical paradigm is to assume that the excitation $E_U(\omega)$ is white noise. This is not quite true for explosives and fricatives in theory, where the power spectrum may not be flat [74], but in practice it provides reasonable approximation.

2.6 AM-FM Effects in Speech

So far we have assumed that speech is perfectly periodic inside a frame, while in fact there exist significant AM-FM effects. There are many sources of AM-FM effects, and two important sources are pitch jitter and amplitude shimmer [1]. Pitch jitter refers to the pitch fluctuations around the steady target that the speaker intends to maintain. Amplitude shimmer refers to amplitude variations that may be due to the time varying characteristics of

vocal tract and vocal fold.

In particular, the special glottal wave patterns would introduce AM-FM effects. These patterns include creaky voice, where only a portion of vocal chords vibrate, causing low and irregular pitch; vocal fry [75, 76], where there is a secondary pulse at the start of the major pulse; and diplophonic [12], where there is a small pulse preceding each major pulse.

With AM-FM effects, equation (2.14) can be rewritten as

$$S_V(\omega) = \sum_d |\alpha_d| \eta_d(t) \exp\left(j\left(d\omega_0(t-\tau) + \angle\alpha_d + d\phi(t)\right)\right) \tag{2.18}$$

where the additional terms $\eta_d(t)$ and $\phi(t)$ refer to the amplitude modulation of the d-th harmonic and phase modulation at the fundamental frequency respectively. Here we impose an important assumption that phase modulation at higher harmonics is proportional to that at the fundamental frequency.

2.7 The Signal Model of PAT

Now we are ready to introduce the signal model of PAT, which is essentially the summary of the above sections.

The proposed PAT model is a frame-based speech model. We introduce the subscript k to denote the DTFTs of the k-th frame. Then, the signal model assumes the speech of frame k, $S_k(\omega)$, can be decomposed in two components - the voiced part and the unvoiced part.

$$\begin{aligned} S_k(\omega) &= S_{Vk}(\omega) + S_{Uk}(\omega) \\ &= \left(a_k E_{Vk}(\omega) + b_k E_{Uk}(\omega)\right) H_k(\omega) \circledast W(\omega) \end{aligned} \tag{2.19}$$

where the second equality is consistent with equation (2.14). $H_k(\omega)$ is the vocal tract transfer function, defined by equation (2.10). $E_{Uk}(\omega)$ is the unvoiced aspiration excitation, which, as discussed, is white Gaussian noise. $E_{Vk}(\omega)$ is the quasi-periodic glottal wave, whose single cycle is defined by equation (2.17), and the resulting amplitude and frequency modulated $S_{Vk}(\omega)$ is given by (2.18). $W(\omega)$ is the frequency response of the rectangular window function.

CHAPTER 3

PROBABILISTIC MODELING OF PAT

The probabilistic model of PAT essentially involves defining the random variables and imposing probabilistic assumptions on the signal model.

3.1 Notation

Now we will introduce some notation that will be used frequently within this section. Denote lower case letters, a, as scalars; lower case bold letters, $\boldsymbol{b}$, as vectors; and upper case bold letters, $\boldsymbol{A}$, as matrices. The terms real$[\cdot]$ and imag$[\cdot]$ denote real and imaginary parts of their argument, respectively, and diag$[\cdot]$ denotes converting the column vector in its argument into a diagonal matrix. The colon in the subscript, $\boldsymbol{a}_{m:n}$, denotes a column vector $[\boldsymbol{a}_m, \boldsymbol{a}_{m+1}, \cdots, \boldsymbol{a}_{n-1}, \boldsymbol{a}_n]^T$.

3.2 The Real DFT Vector Form

To facilitate probabilistic representation using vectors and matrices, we will switch from the DTFT domain to the DFT domain, with real and imaginary parts separated. Specifically, for some DTFT $X(\omega)$, denote its real DFT vector $\boldsymbol{x}$ as

$$\boldsymbol{x} = \sqrt{\frac{2}{T}}\Bigg[\frac{1}{\sqrt{2}}X(0), \text{real}\left[X\left(\frac{2\pi}{T}\right), X\left(\frac{4\pi}{T}\right), \cdots, X\left(\frac{(T-2)\pi}{T}\right)\right] \\ \frac{1}{\sqrt{2}}X\left(\pi\right), \text{imag}\left[X\left(\frac{2\pi}{T}\right), X\left(\frac{4\pi}{T}\right), \cdots, X\left(\frac{(T-2)\pi}{T}\right)\right]\Bigg]^T \tag{3.1}$$

where T is the frame length, and also the length of the real DFT vector. The reason we define the real DFT vector this way is to preserve Parseval's

theorem [67], i.e. the real DFT vector norm is equal to the time-domain vector norm. Before further mathematical details are given, here are some intuitions. If the time-domain signal is real, then $X(\omega)$ is conjugate symmetric with respect to π. So the DFT between 0 and π is sufficient to represent and recover the whole DFT; that is why the real DFT vector only contains frequency points between 0 and π. Also, under conjugate symmetry, $X(0)$ and $X(\pi)$ must be real, so no imaginary parts of $X(0)$ and $X(\pi)$ are included in the real DFT vector.

Now define the time-domain vector as

$$\boldsymbol{x}_{\text{time}} = [x(0), x(1), \cdots, x(T-1)]^T \tag{3.2}$$

and define the real DFT transform matrix as

$$\boldsymbol{D} = \sqrt{\frac{2}{T}} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \cdots & \frac{1}{\sqrt{2}} \\ 1 & \cos\left(\frac{2\pi}{T}\right) & \cos\left(\frac{4\pi}{T}\right) & \cdots & \cos\left(\frac{2(T-1)\pi}{T}\right) \\ 1 & \cos\left(\frac{4\pi}{T}\right) & \cos\left(\frac{8\pi}{T}\right) & \cdots & \cos\left(\frac{4(T-1)\pi}{T}\right) \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \cos\left(\frac{(T-2)\pi}{T}\right) & \cos\left(\frac{2(T-2)\pi}{T}\right) & \cdots & \cos\left(\frac{(T-1)(T-2)\pi}{T}\right) \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \cdots & -\frac{1}{\sqrt{2}} \\ 0 & \sin\left(\frac{2\pi}{T}\right) & \sin\left(\frac{4\pi}{T}\right) & \cdots & \sin\left(\frac{2(T-1)\pi}{T}\right) \\ 0 & \sin\left(\frac{4\pi}{T}\right) & \sin\left(\frac{8\pi}{T}\right) & \cdots & \sin\left(\frac{4(T-1)\pi}{T}\right) \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \sin\left(\frac{(T-2)\pi}{T}\right) & \sin\left(\frac{2(T-2)\pi}{T}\right) & \cdots & \sin\left(\frac{(T-1)(T-2)\pi}{T}\right) \end{bmatrix} \tag{3.3}$$

It can be shown that

$$\boldsymbol{x} = \boldsymbol{D}\boldsymbol{x}_{\text{time}} \tag{3.4}$$

by noticing that $\boldsymbol{D}$ is simply rearranging and scaling the real and imaginary parts of each row of the DFT matrix the same way equation (3.1) rearranges and scales the true DFT.

Also, it can be shown that $\boldsymbol{D}$ is orthonormal, i.e.

$$\boldsymbol{D}\boldsymbol{D}^T = \boldsymbol{D}^T\boldsymbol{D} = \boldsymbol{I} \tag{3.5}$$

and therefore Parseval's theorem holds for the real DFT vector, i.e.

$$\boldsymbol{x}^T\boldsymbol{x} = \boldsymbol{x}_{\text{time}}^T\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{x}_{\text{time}} = \boldsymbol{x}_{\text{time}}^T\boldsymbol{x}_{\text{time}} \tag{3.6}$$

3.3 Model Overview

Denote $\boldsymbol{s}_k$, $\boldsymbol{s}_{Vk}$ and $\boldsymbol{s}_{Uk}$ as the real DFT vector of $S_k(\omega)$, $S_{Vk}(\omega)$ and $S_{Uk}(\omega)$ respectively. Then the observed variable in frame k is $\boldsymbol{s}_k$.

Denote $\boldsymbol{z}_k$ as a set of hidden variables, whose elements will be defined soon. Then the probabilistic model is defined as a hidden Markov model.

$$p(\{\boldsymbol{s}_k, \boldsymbol{z}_k\}) = \prod_{k=1}^{K} p(\boldsymbol{s}_k|\boldsymbol{z}_k)p(\boldsymbol{z}_k|\boldsymbol{z}_{k-1}) \tag{3.7}$$

where $p(\boldsymbol{z}_1|\boldsymbol{z}_0)$ denotes $p(\boldsymbol{z}_1)$ for notational simplicity. So the probabilistic model boils down to defining $p(\boldsymbol{s}_k|\boldsymbol{z}_k)$ and $p(\boldsymbol{z}_k|\boldsymbol{z}_{k-1})$.

Then according to equation (2.19),

$$\boldsymbol{s}_k = \boldsymbol{s}_{Vk} + \boldsymbol{s}_{Uk} \tag{3.8}$$

and therefore $p(\boldsymbol{s}_k|\boldsymbol{z}_k)$ can be determined by $p(\boldsymbol{s}_{Vk}|\boldsymbol{z}_k)$ and $p(\boldsymbol{s}_{Uk}|\boldsymbol{z}_k)$.

Section 3.4 defines $p(\boldsymbol{s}_{Uk}|\boldsymbol{z}_k)$; section 3.5 defines $p(\boldsymbol{s}_{Vk}|\boldsymbol{z}_k)$; section 3.6 defines $p(\boldsymbol{z}_k|\boldsymbol{z}_{k-1})$.

3.4 The Unvoiced Model

Denote $\boldsymbol{e}_{Uk}$ and $\boldsymbol{h}_k$ as the real DFT vectors of $E_{Ut}(\omega)$ and $H_t(\omega)$ respectively. According to (2.19),

$$\boldsymbol{s}_{Uk} = b_k \text{diag}\,(\boldsymbol{h}_k)\,\boldsymbol{e}_{Uk} \tag{3.9}$$

where b_k is one hidden variable. $\boldsymbol{h}_k$ can be completely determined by the complex cepstral coefficients at positive low quefrencies, i.e.

$$\hat{\boldsymbol{h}}_k = [h[0], \cdots, h[C]]^T \tag{3.10}$$

according to equation (2.11), and thus $\hat{\boldsymbol{h}}_k$ is a hidden variable. The windowing function can be omitted because it is a rectangular window. Since the time-domain signal of $\boldsymbol{e}_{Uk}$ is white Gaussian noise and the real DFT transform $\boldsymbol{D}$ is orthogonal, $\boldsymbol{e}_{Uk}$ is also white Gaussian noise, i.e.

$$\boldsymbol{e}_{Uk} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{3.11}$$

Combining equations (3.9) and (3.11) we have

$$p(\boldsymbol{s}_{Uk}|\boldsymbol{z}_k) = \mathcal{N}(\boldsymbol{s}_{Uk}; \boldsymbol{0}, b_k^2 \mathrm{diag}(\boldsymbol{h}_k)^2) \tag{3.12}$$

3.5 The Voiced Model

For notational ease, the frame subscript k will be omitted throughout this section, when there is no ambiguity introduced.

3.5.1 Adapted Bayesian Spectral Estimation Model

The major randomness in $\boldsymbol{s}_{Vk}$ lies in its AM-FM effect. Formally, rewrite equation (2.18) as

$$s_v(t) = \sum_d \boldsymbol{x}_d(t)^T \boldsymbol{\xi}_d(t) \tag{3.13}$$

where

$$\boldsymbol{x}_d(t) = \begin{bmatrix} |\alpha_d| \cos(d\omega_0(t-\tau) + \angle\alpha_d) \\ |\alpha_d| \sin(d\omega_0(t-\tau) + \angle\alpha_d) \end{bmatrix} \tag{3.14}$$

which is essentially the vector form of the clean signal, and

$$\boldsymbol{\xi}_d(t) = \begin{bmatrix} \eta_d(t) \cos(d\phi(t)) \\ \eta_d(t) \sin(d\phi(t)) \end{bmatrix} \tag{3.15}$$

which is essentially the vector form of the AM-FM random variations.

In Bayesian spectral estimation (BSE)[57], if $d\phi(t)$ is uniformly distributed in $[-\pi, \pi]$, $\boldsymbol{\xi}_d(t)$ can be modeled as a multivariate Gaussian with 0 mean and diagonal identity covariance matrix. However, in PAT, the uniform distribution of $d\phi(t)$ is not a reasonable assumption. Nevertheless, $\boldsymbol{\xi}_d(t)$ can still be reasonably approximated by a joint Gaussian with matched first and

second moments, as will be shown in the next subsection.

3.5.2 The Model of $\boldsymbol{\xi}_d(t)$

Before deriving the appropriate moments for $\boldsymbol{\xi}_d(t)$, we first state our assumptions on $\eta_d(t)$ and $\phi(t)$:

- The distribution of $\eta_d(t)$ is symmetric and centered at 0.
- $\phi(t)$ is small with respect to π, and has symmetric and unimodal distribution centered at 0.

With these assumptions, it can be shown that $\xi_d^{(1)}$ and $\xi_d^{(2)}$, the two elements of $\boldsymbol{\xi}_d$, are uncorrelated and both 0 mean, which can then be reasonably assumed to satisfy independent Gaussian distribution:

$$\boldsymbol{\xi}_d(t) \sim \mathcal{N}\left(\mathbf{0}, \sigma_\xi^2 \begin{bmatrix} 1 & 0 \\ 0 & \rho_d^2 \end{bmatrix}\right) \tag{3.16}$$

where ρ_d is the ratio of their standard deviation. The reason why ρ_d depends on d is that $d\phi(t)$ depends linearly on d. Intuitively, if d is very small, $d\phi(t)$ is close to 0. From (3.15), the variance of $\xi_d^{(2)}$ is close to 0, and therefore ρ_d is close to 0. On the other hand, if d goes to infinity, $d\phi(t)$ will approach a uniform distribution, and therefore ρ_d will approach 1 (the model becomes the standard BSE).

Formally, notice that by (3.15)

$$\mathbb{A}_{[0,\pi]} d\phi(t) = \arctan\left(\frac{\xi_d^{(2)}(t)}{\xi_d^{(1)}(t)}\right) \stackrel{\text{def}}{=} \psi_d(t) \tag{3.17}$$

where $\mathbb{A}_{[0,\pi]}$ denotes the principal value within the interval $[0, \pi]$. We need to find a distribution of unaliased $d\phi(t)$ such that its aliased distribution, i.e. the distribution of $\mathbb{A}_{[0,\pi]} d\phi(t)$, fits that of $\psi_d(t)$ under the Gaussianity assumption in (3.16).

Now we will find such a distribution. The marginal distribution of $\psi_d(t)$ is given by

$$p(\psi_d(t) = \theta) = \frac{\rho_d}{\rho_d^2 + (1 - \rho_d^2)\sin^2\theta} \cdot \frac{1}{2\pi}, \theta \in [-\pi, \pi] \tag{3.18}$$

Rewrite equation (3.18) as

$$p(\psi_d(t) = \theta) \propto \frac{1}{[1 - f(\rho_d)e^{2j\theta}]\,[1 - f(\rho_d)e^{-2j\theta}]}, \theta \in [-\pi, \pi] \tag{3.19}$$

where

$$f(\rho_d) = \sqrt{\frac{1+\rho_d}{1-\rho_d}} \tag{3.20}$$

Now we can borrow the correspondence between Fourier transform (FT) and DTFT to find an unaliased distribution. Note that DTFT is the aliased version of FT within the interval $[0, f_s]$, where f_s is the sampling frequency. For right-sided exponential signals, we have the following correspondence:

$$\frac{1}{a + j\omega} \leftrightarrow \frac{1}{1 - \alpha e^{-j\omega}} \tag{3.21}$$

where the left-hand side is FT, and the right-hand side is DTFT; a and α are related by time-domain sampling:

$$\alpha = \exp\left(-\frac{a}{f_s}\right) \text{ or } a = -f_s \log \alpha \tag{3.22}$$

Plugging equation (3.21) into equation (3.19), we can see that one of the unaliased distributions is a Cauchy distribution, namely

$$p_{d\phi(t)}(\varphi) = \frac{1}{\pi\gamma_d} \cdot \frac{\gamma_d^2}{\varphi^2 + \gamma_d^2} \tag{3.23}$$

where the parameter γ_d satisfies

$$\gamma_d = -f_s \log\left(\sqrt{\frac{1+\rho_d}{1-\rho_d}}\right) \tag{3.24}$$

Cauchy distribution has a scaling property: if $\phi(t) \sim \text{Cauchy}(\gamma_1)$, then $d\phi(t) \sim \text{Cauchy}(d\gamma_1)$. Therefore

$$\gamma_d = d\gamma_1 \stackrel{\text{def}}{=} d\omega_0\gamma$$

By (3.24), we have

$$\rho_d = \tanh(2d\omega_0\gamma) \tag{3.25}$$

which agrees with our intuitive notion of its asymptotic behavior as d varies.

Equations (3.16) and (3.25) characterize the model of $\boldsymbol{\xi}_d(t)$.

3.5.3 Relation of $\boldsymbol{\xi}_d(t)$ across t

Since vocal tract movement, glottal fold variation and pitch variation are slowly time varying, $\boldsymbol{\xi}_d(t)$ should also be slowly time varying. BSE proposed a solution which is adopted in PAT: $\boldsymbol{\xi}_d(t)$'s are modeled as a first-order autoregressive process.

$$\boldsymbol{\xi}_d(t) = \lambda_d \boldsymbol{\xi}_d(t-1) + \boldsymbol{\varepsilon}_d(t) \tag{3.26}$$

where

$$\boldsymbol{\varepsilon}_d(t) \sim \mathcal{N}\left(\mathbf{0}, \sigma_\varepsilon^2 \begin{bmatrix} 1 & 0 \\ 0 & \rho_d^2 \end{bmatrix}\right) \tag{3.27}$$

and is independent of $\boldsymbol{\xi}(t-1)$.

By quasi-stationarity of speech, it is reasonable to assume that the autoregressive process in (3.26) is close to a stationary distribution. It can be shown that a stationary distribution of $\boldsymbol{\xi}(t)$ implies (3.16), with σ_ξ determined by

$$\sigma_\xi = \frac{\sigma_\varepsilon}{\sqrt{1-\lambda_d^2}} \tag{3.28}$$

where λ_d is the first-order autoregressive coefficient. Again, the reason why λ_d depends on d is that $d\phi(t)$ depends linearly on d. Intuitively, as d goes up, the AM/FM variation becomes larger, and therefore $\lambda_d(d)$ becomes closer to 0.

It is generally hard to determine the relationship between λ_d and d. With some approximating assumptions [48], we could approximate that λ_d decreases exponentially with d, i.e.

$$\lambda_d = \exp(-d\delta) \tag{3.29}$$

where δ is the parameter of the Cauchy-distributed increment of $\phi(t)$. Equation (3.29) agrees with our intuitive notion of its asymptotic behavior.

3.5.4 The PDF of $\boldsymbol{s}_{Vk}$

With the above derivation, we are ready to summarize the PDF of $\boldsymbol{s}_{Vk}$ conditional on $\boldsymbol{z}_k$. Since $\boldsymbol{\xi}_d(t)$ is also a zero-mean Gaussian process and $\boldsymbol{s}_{Vk}$ is its linear transformation, $\boldsymbol{s}_{Vk}$ is a zero-mean multicovariate Gaussian, whose distribution is determined once its second moment is specified. Now we will derive its second moment.

From equation (3.26), we know that

$$\mathbb{E}\left(\boldsymbol{\xi}_d(t)\boldsymbol{\xi}_d(t-t')^T\right) = \lambda_d^{2t'}\sigma_\varepsilon^2 \begin{bmatrix} 1 & 0 \\ 0 & \rho_d^2 \end{bmatrix} \tag{3.30}$$

From (3.13), we can obtain the autocorrelation function of the time domain signal $s_V(t)$

$$\begin{aligned} &R_{s_V}(t') \stackrel{\text{def}}{=} \mathbb{E}\left(s_V(t)s_V(t-t')\right) \\ &= \boldsymbol{x}_d(t)^T\mathbb{E}\left(\boldsymbol{\xi}_d(t)\boldsymbol{\xi}_d(t-t')^T\right)\boldsymbol{x}_d(t-t') \\ &= |\alpha_d|^2\lambda_d^{2t'}\sigma_\varepsilon^2\big(\cos\left(d\omega_0(t-\tau)+\angle\alpha_d\right)\cos\left(d\omega_0(t-t'-\tau)+\angle\alpha_d\right) \\ &+ \rho_d^2\sin\left(d\omega_0(t-\tau)+\angle\alpha_d\right)\sin\left(d\omega_0(t-t'-\tau)+\angle\alpha_d\right)\big) \end{aligned} \tag{3.31}$$

Hence, the distribution of $\boldsymbol{s}_{Vk}$, the real DFT vector of $s_V(t)$, is given by

$$p(\boldsymbol{s}_{Vk}|\boldsymbol{z}_k) = \mathcal{N}\left(\boldsymbol{s}_{Vk}; \boldsymbol{0}, \boldsymbol{D}\boldsymbol{R}_{s_V}\boldsymbol{D}^T\right) \tag{3.32}$$

where $\boldsymbol{R}_{s_V}$ is the autocorrelation matrix of $s_V(t)$, which is a Toeplitz matrix whose t'-th subdiagonal elements are $R_{s_V}(t')$. The hidden variables include all the variables that determine the signal vector $\boldsymbol{x}_d(t)$, i.e. a_k in equation (2.19), $g_{1k}, \beta_{1k}, g_{2k}$ in equation (2.17), $\hat{\boldsymbol{h}}_k$ as specified in section 3.4, ω_{0k} and τ_k as in equation (3.14). As a reminder, subscript k is added to distinguish the hidden variables in different frames.

3.6 Hidden Variables Transitions

As a summary, all the hidden variables are given as follows:

$$\boldsymbol{z}_k = \left[a_k, b_k, g_{1k}, \beta_{1k}, g_{2k}, \hat{\boldsymbol{h}}_k^T, \omega_{0k}, \tau_k\right]^T \tag{3.33}$$

Since articulators and pitch are slowly time varying, z_k tends to transit smoothly among frames. Therefore, we apply a random walk to model this behavior:

$$p(\boldsymbol{z}_k|\boldsymbol{z}_{k-1}) = \mathcal{N}\left(\boldsymbol{z}_k; \boldsymbol{z}_{k-1}, \text{diag}(\boldsymbol{\sigma}_z^2)\right) \tag{3.34}$$

where $\boldsymbol{\sigma}_z^2$ is the variance of the innovation of each dimension of $\boldsymbol{z}_k$. Notice that the last dimension, i.e. the innovation variance of τ_k, is set to ∞ because τ_k does not transit smoothly. This is equivalent to imposing a non-informative transition prior on τ_k.

3.7 Model Summary

To sum up, the observed variable of PAT is $\boldsymbol{s}_k$, and the hidden variables $\boldsymbol{z}_k$ are defined in (3.33). The joint probability of all the variables is given in (3.7), where according to equations (3.8), (3.12) and (3.32),

$$p(\boldsymbol{s}_k|\boldsymbol{z}_k) = \mathcal{N}\left(\boldsymbol{s}_k; \boldsymbol{0}, \boldsymbol{D}\boldsymbol{R}_{s_V}\boldsymbol{D}^T + b_k^2\text{diag}(\boldsymbol{h}_k)^2\right) \tag{3.35}$$

and $p(\boldsymbol{z}_k|\boldsymbol{z}_{k-1})$ is given in equation (3.34). The model parameters include

$$\Theta = \left\{\gamma, \delta, \boldsymbol{\sigma}_z^2\right\} \tag{3.36}$$

3.8 Model Inference

The task of model inference is to infer the value of hidden variables $\{\boldsymbol{z}_k\}$ given the observed $\{\boldsymbol{s}_k\}$. To reduce computational complexity, we adopt the online MAP criteria:

$$\begin{aligned} \hat{\boldsymbol{z}}_k =& \underset{\boldsymbol{z}_k}{\text{argmax}}\ p\left(\boldsymbol{z}_k|\boldsymbol{s}_{1:k}, \boldsymbol{z}_{1:k-1} = \hat{\boldsymbol{z}}_{1:k-1}\right) \\ =& \underset{\boldsymbol{z}_k}{\text{argmax}}\ p(\boldsymbol{z}_k|\boldsymbol{z}_{k-1} = \hat{\boldsymbol{z}}_{k-1})p(\boldsymbol{s}_k|\boldsymbol{z}_k) \end{aligned} \tag{3.37}$$

We use gradient ascent to solve the optimization problem. To avoid getting trapped in local optima, specifically for ω_{0k} and τ_k, we have special initialization schemes for them.

For ω_{0k}, we incorporate the information in the autocorrelation function

$R_{Sv}(t')$ as defined in equation (3.31). It is well-known that for quasi-periodic signal, $R_{Sv}(t')$ would have peaks at multiples of the period. Therefore, the initial values of ω_{0k} are set such that their corresponding pitch period lies at the highest peaks of $R_{Sv}(t')$. In practice, we choose the 5 highest peaks to avoid double and half pitch errors that are commonly encountered in autocorrelation-based pitch tracking algorithms.

For τ_k, we incorporate the information in the short-time energy function $e(t)$, which is defined as

$$e(t) = \sum_{t'=t-r}^{t+r} s(t')^2 \tag{3.38}$$

It has been shown that GCIs occur where the glottal opening is maximum, and therefore the short-time energy reaches local maxima. So the initial values of τ are set to the 5 highest peaks of the short-time energy function. Unlike the case with pitch, where different autocorrelation peaks correspond to different pitches, two values of τ are equivalent if they differ by multiples of the pitch period, and thus the initialization of τ is less sensitive to picking a wrong peak.

There are a total of 25 (5 for ω_{0k} and 5 for τ_k) different initialization combinations, and therefore the optimization would run 25 times for each frame, and the local optimum with the highest posterior probability is chosen as the inferred values of the hidden variables.

CHAPTER 4

EXPERIMENTS

This chapter presents some experiment results that demonstrate the potential of PAT in various speech processing tasks. The experiments will demonstrate the capability of the PAT model of inferring the hidden variables and reconstructing speech, including the phase spectrum.

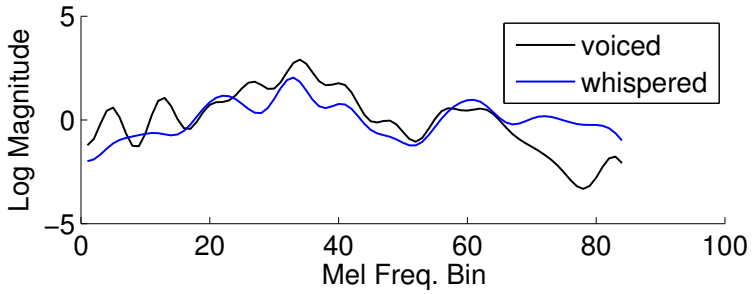
4.1 Configuration

Except for the experiment introduced in section 4.2, all the experiments are performed on the Edinburgh speech corpus [3]. The sampling rate is 10 kHz. Speech is segmented into 30 ms frames with 10 ms frame shift. All the figures demonstrated are from speaker 1, utterance 1. The dimension of $\hat{h}(\hat{t})$ is set to 26.

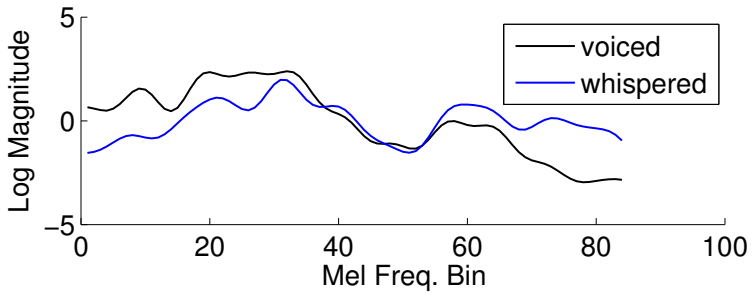
4.2 The "Glottal Free" Vocal Tract Estimate

According to chapter 1, current vocal tract representations such as LPC and MFCC essentially mix glottal wave and vocal tract transfer function, and their separation cannot be obtained without a unified model like PAT. Therefore, PAT provides some insights into disentangled vocal tract. To illustrate this, 2 extreme utterances of /ah/ are recorded, one uttered with voiced excitation and the other whispered. The idea is that the vocal tract shapes in both cases are similar, but according to section 2.5, one has spectral tilt and the other does not. It is expected that the PAT model would give more consistent estimates of the vocal tract of the two cases than MFCC does.

Figure 4.1 compares the mean of the envelope estimates (the estimate of $\boldsymbol{h}_k$) of both cases by the two methods. It turns out that both MFCC and



(a) PAT estimation



(b) MFCC estimation

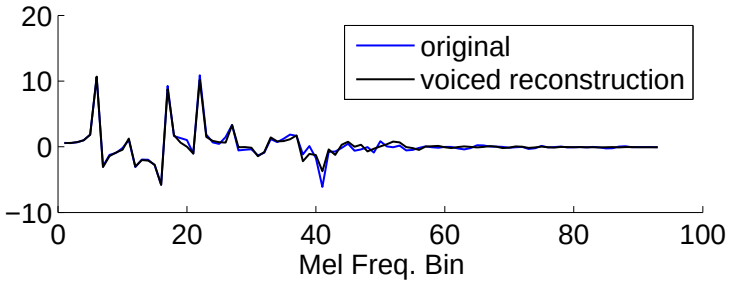
Figure 4.1: The means of the estimated vocal tract frequency response / spectral envelope for a voiced-excited and a whispered utterance of /ah/

PAT have almost the same envelope estimates for the whispered case, but very different for voiced. PAT has much more consistent estimates for both cases, especially in the mid-frequency. The norm of the differences between the means of the estimates for the two cases is 10.93 for PAT, as opposed to 13.15 for MFCC.

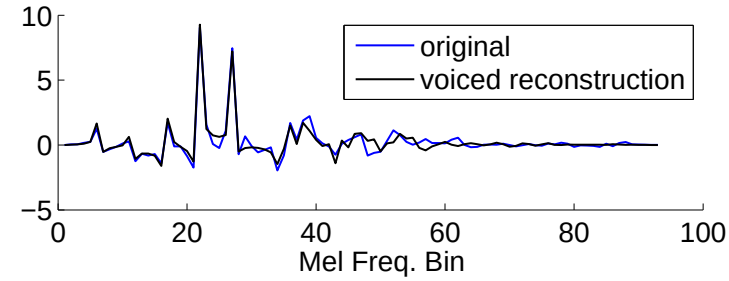
4.3 Phase Reconstruction

As mentioned in chapter 1, the synthesis using parameters estimated separately does not necessarily resemble original speech. The second experiment shows that PAT is able to yield parameter estimates that are accurate for synthesis.

Figure 4.2 compares both real and imaginary spectra of the voiced speech frame taken from the Edinburgh speech corpus reconstructed by PAT parameter estimates (namely $\boldsymbol{x}_k$ in equation 3.14) and those of the original speech. We can see that the reconstruction almost overlaps with the original in low frequencies in both spectra, which shows that PAT models speech very



(a) Real spectrum



(b) Imaginary spectrum

Figure 4.2: Comparison of real and imaginary spectrum for a voiced speech frame.

accurately.

4.4 Reconstruction of Speech with Heavy AM/FM Effect

This section demonstrates the effect of AM/FM modeling. Speech frames with significant AM/FM effect are studied. Figures 4.3 and 4.4 display the reconstructed magnitude spectrum $\boldsymbol{s}_{Vk}$ of some speech frames. The left panel is reconstructed by the PAT with AM/FM modeling, the right by the PAT without. The black solid line is the reconstructed magnitude spectrum, the blue line is the original magnitude spectrum, and the dashed line is the estimated spectral envelope. Our first observation is that in the original magnitude spectrum, the bandwidth of the pitch pulses is small when the frequency is low, and increases as frequency goes up. This widening of pitch pulses is the major effect of AM/FM, and it becomes more significant in mid and high frequencies, which agrees with (3.25) and (3.29).

As for reconstruction accuracy, the PAT significantly underestimates voiced

energy in mid and high frequencies. This is because without AM/FM modeling, the PAT does not account for the widening of the pitch pulses, and ascribes this variation to unvoiced energy. On the other hand, the AM/FM model is able to more accurately estimate the spectral envelope.

4.5 GCI Location

GCI estimation is indicative of PAT's ability in phase modeling and pitch tracking. According to chapter 2, τ_k is the delay of the first GCI relative to the beginning of frame k. Also, we know that GCI are periodic at the fundamental frequency. Estimated GCI locations of frame t are thus $\tau_k + 2m\pi/\omega_{0k}$, where m is nonnegative integer. Since GCIs of different frames are estimated separately, we can judge the accuracy by checking: 1) if GCIs of different frames are consistent, i.e. if they form a quasi-periodic sequence; 2) if they appear at the energy bursts of the original speech.

Figure 4.5 plots GCI locations as impulses against original speech waveform. As can be seen, GCIs, around 3 or 4 instances in each frame, form a quasi-periodic signal with rare exceptions. What is more, they tend to appear consistently at the largest negative to positive jump within a period in the original speech wave, where short-time energy is generally greatest. This result shows that PAT can control well for group delay and pitch, and thus achieves similar performance to pitch-synchronous analysis.

4.6 Pitch Tracking

Pitch tracking by PAT is essentially the inference of $f_{0,n}$. Since a U/V decision scheme for PAT has yet to be developed, we extract pitch on labeled voiced segments only, and compare against a pitch-tracking benchmark, GetF0 [4]. Both algorithms are run over the complete Edinburgh dataset. For fair comparison, we compare the pitch tracking results of all the voiced frames that are also correctly classified as voiced by GetF0, in terms of the following 2 criteria:

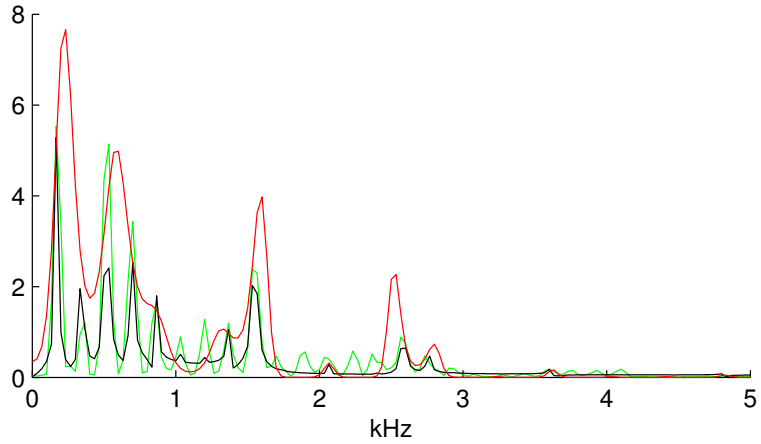
Gross Pitch Error (GPE): The percentage of frames whose pitch estimates deviate from ground truth by more than 20%.

Table 4.1: Pitch tracking results on Edinburgh dataset

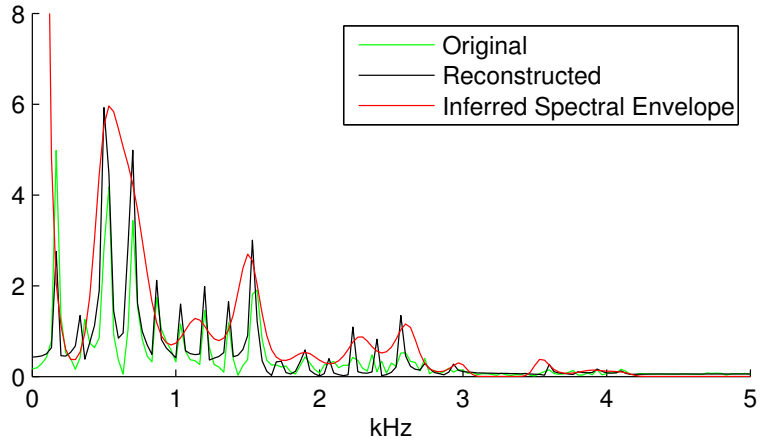
	PAT	**GetF0**
GPE (%)	2.10	2.07
RMS (Hz)	5.052	5.780

Root Mean Squared Error (RMS): The averaged mean squared error in Hz over the frames free of GPE.

Table 4.1 shows the results. As can be seen, PAT has GPE level comparable to that of GetF0, but much smaller RMS, which means PAT inference is more accurate.

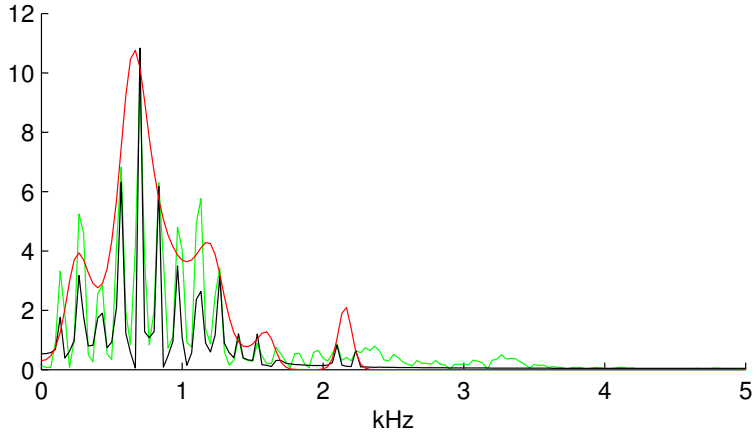


(a) Frame 32, without AM/FM modeling

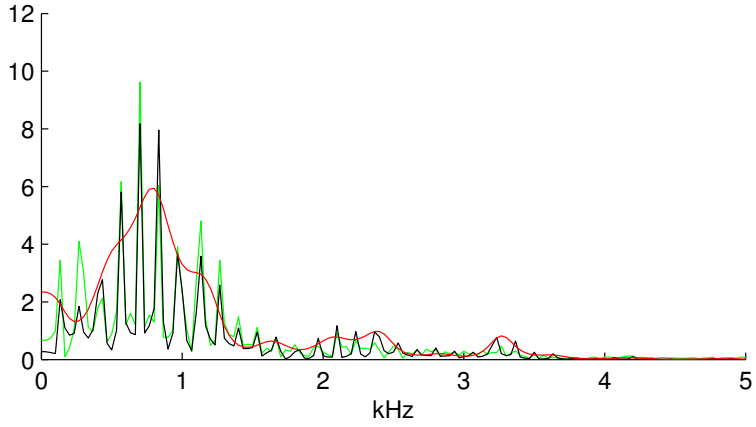


(b) Frame 32, with AM/FM modeling

Figure 4.3: Magnitude spectrum of reconstructed voiced speech (black line) against original magnitude spectrum (blue line) for frame 32. AM/FM modeling (right panels) is able to reclaim much of the voiced energy overlooked by the model without AM/FM modeling (left panels), especially in mid-frequencies.



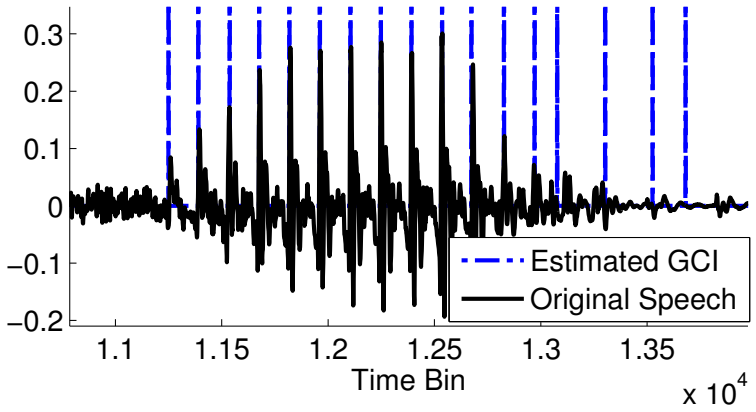
(a) Frame 60, without AM/FM modeling



(b) Frame 60, with AM/FM modeling

Figure 4.4: Magnitude spectrum of reconstructed voiced speech (black line) against original magnitude spectrum (blue line) for frame 60.

Figure 4.5: Estimation of GCI location of the utterance 'park'.



CHAPTER 5

CONCLUSION AND FUTURE DIRECTIONS

This thesis proposes a probabilistic generative model of speech, which jointly models pitch, glottal wave, aspiration, vocal tract and AM/FM effect. The PAT model applies the well-founded theories on speech signal processing and imposes theoretically reasonable probabilistic assumptions among the variables. A couple of experiments demonstrate that the inference of the hidden variables is accurate and the joint modeling is able to control for the interference induced by the variations of other variables.

Here we discuss some future directions. First, many probabilistic assumptions are trade-offs between theoretical plausibility and inference tractability. Sometimes the model accuracy is compromised for a more tractable inference scheme. For example, the Gaussian assumption on the AM/FM effect makes the conditional probability a closed-form expression by approximating the smooth transition of AM and FM components, i.e. $\eta_d(t)$ and $\phi(t)$ as in equation (2.18), to that of $\boldsymbol{\xi}_d(t)$ as in equation (3.15). However, according to [55], this approximation suffers from significant error when $\eta_d(t)$ is small. As another example, the three-pole model of glottal wave is a simplification of the LF model by reducing the number of parameters, and thus computational complexity. Yet the approximation error is non-negligible. Therefore, we would like to find a better inference scheme so that some of these approximations can be eliminated. The Monte-Carlo based approaches [77] have gained popularity to evaluate complex distributions, and can be potentially applied to the PAT model.

Second, in order for the PAT model to be applied to more sophisticated speech processing tasks, such as speech enhancement and source separation, it should be adapted to accommodate interference and noise. Currently the PAT model only considers perfectly clean speech, and assumes that all the variations are due to variations of speech signal. To incorporate environment noise, both the probabilistic assumptions and the inference algorithm should

be adapted.

To sum up, our ultimate goal is to develop a probabilistic acoustic model for speech, which accurately defines the probabilistic space spanned by speech, and can be applied to speech enhancement, source separation, pitch tracking and speech recognition with improved performance and efficiency.

REFERENCES

[1] T. F. Quatieri, *Discrete-Time Speech Signal Processing*. Pearson Education India, 2002.

[2] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[3] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching." in *Proc. Eurospeech*. International Speech Communication Association, 1993.

[4] D. Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech Coding and Synthesis*, vol. 495, p. 518, 1995.

[5] B. S. Lee, "Noise robust pitch tracking by subband autocorrelation classification," Ph.D. dissertation, Columbia University, 2012.

[6] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 3, pp. 229–241, 2003.

[7] K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–361–I–364.

[8] M. M. Sondhi, "New methods of pitch extraction," *Audio and Electroacoustics, IEEE Transactions on*, vol. 16, no. 2, pp. 262–266, 1968.

[9] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[10] H. Kameoka, N. Ono, and S. Sagayama, "Speech spectrum modeling for joint estimation of spectral envelope and fundamental frequency," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1507–1516, 2010.

[11] T. A. Stephenson, M. M. Doss, and H. Bourlard, "Speech recognition with auxiliary information," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 3, pp. 189–203, 2004.

[12] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, p. 820, 1990.

[13] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, "Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis," *Speech Communication*, vol. 55, no. 2, pp. 278–294, 2013.

[14] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Acoustics, Speech and Signal Processing, 2008.(ICASSP 2008). IEEE International Conference on*. IEEE, 2008, pp. 3933–3936.

[15] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," *Sadhana*, vol. 36, no. 5, pp. 713–727, 2011.

[16] K. Achan, S. Roweis, A. Hertzmann, and B. Frey, "A segment based probabilistic generative model of speech," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP 2005). IEEE International Conference on*, vol. 5. IEEE, 2005, pp. v–221.

[17] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 982–994, 2007.

[18] P. J. Jackson and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 7, pp. 713–726, 2001.

[19] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech communication*, vol. 11, no. 2, pp. 109–118, 1992.

[20] T. Drugman, B. Bozkurt, and T. Dutoit, "Causal–anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Communication*, vol. 53, no. 6, pp. 855–866, 2011.

[21] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1066–1074, 2007.

[22] A. Cichocki, R. Zdunek, and S.-i. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 5. IEEE, 2006, pp. V–V.

[23] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 550–563, 2010.

[24] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. John Wiley & Sons, 2009.

[25] S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 129–132.

[26] T.-W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *Signal Processing Letters, IEEE*, vol. 6, no. 4, pp. 87–90, 1999.

[27] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. J. McKeown, V. Iragui, and T. J. Sejnowski, "Removing electroencephalographic artifacts by blind source separation," *Psychophysiology*, vol. 37, no. 02, pp. 163–178, 2000.

[28] M. E. Davies and C. J. James, "Source separation using single channel ICA," *Signal Processing*, vol. 87, no. 8, pp. 1819–1832, 2007.

[29] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 2, pp. 666–678, 2006.

[30] N. Correa, T. Adalı, and V. D. Calhoun, "Performance of blind source separation algorithms for fMRI analysis using a group ICA method," *Magnetic Resonance Imaging*, vol. 25, no. 5, pp. 684–694, 2007.

[31] S. S. Tirumala and M. I. Mandel, "Exciting estimated clean spcectra for speech resynthesis," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2015.

[32] M. Mandel, Y. S. Cho, and Y. Wang, "Learning a concatenative resynthesis system for noise suppression," in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on.* IEEE, 2014, pp. 582–586.

[33] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 583–590, 2005.

[34] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 153–165, 2011.

[35] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.* IEEE, 2011, pp. 4564–4567.

[36] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modeling," in *6th ISCA Speech Synthesis Workshop*, 2007, pp. 131–136.

[37] K. Sang-Jin and H. Minsoo, "Two-band excitation for HMM-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 1, pp. 378–381, 2007.

[38] J. P. Cabral, S. Renals, J. Yamagishi, and K. Richmond, "HMM-based speech synthesiser using the LF-model of the glottal source," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.* IEEE, 2011, pp. 4704–4707.

[39] G. Fant, J. Liljencrants, and Q.-g. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[40] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *Signal Processing, IEEE transactions on*, vol. 52, no. 7, pp. 1830–1847, 2004.

[41] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on.* IEEE, 2014, pp. 1562–1566.

[42] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Proceedings of the annual conference of the International Speech Communication Association (INTERSPEECH)*, 2014.

[43] S. T. Roweis, "One microphone source separation," in *NIPS*, vol. 13, 2000, pp. 793–799.

[44] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 191–199, 2006.

[45] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[46] Z. Ou and Y. Zhang, "Probabilistic acoustic tube: a probabilistic generative model of speech for speech analysis/synthesis," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012, pp. 841–849.

[47] Y. Zhang, Z. Ou, and M. Hasegawa-Johnson, "Improvement of probabilistic acoustic tube model for speech decomposition," in *Acoustics, Speech, and Signal Processing, 2014. Proceedings.(ICASSP 2014). IEEE International Conference on.* IEEE, 2014.

[48] Y. Zhang, Z. Ou, and M. Hasegawa-Johnson, "Incorporating AM-FM effect in voiced speech for probabilistic acoustic tube model," in *International Conference on Artificial Intelligence and Statistics*, 2012, pp. 841–849.

[49] B. J. Frey and N. Jojic, "Estimating mixture models of images and inferring spatial transformations using the EM algorithm," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 1. IEEE, 1999.

[50] K. Steiglitz and B. Dickinson, "Phase unwrapping by factorization," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 30, no. 6, pp. 984–991, 1982.

[51] B. Bozkurt and T. Dutoit, "Mixed-phase speech modeling and formant estimation, using differential phase spectrums," in *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003.

[52] B. Doval, C. d'Alessandro, and N. Henrich, "The voice source as a causal/anticausal linear filter," in *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003.

[53] A. Potamianos and P. Maragos, "A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation," *Signal Processing*, vol. 37, no. 1, pp. 95–120, 1994.

[54] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 4, pp. 744–754, 1986.

[55] R. Turner and M. Sahani, "Probabilistic amplitude and frequency demodulation," in *Advances in Neural Information Processing Systems*, 2011, pp. 981–989.

[56] R. E. Turner and M. Sahani, "Decomposing signals into a sum of amplitude and frequency modulated sinusoids using probabilistic inference," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 2173–2176.

[57] Y. Qi, T. P. Minka, and R. W. Picara, "Bayesian spectrum estimation of unevenly sampled nonstationary data," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2002, pp. II–1473.

[58] L. L. Beranek, *Acoustics*. Acoustical Society of America, 1996.

[59] P. M. Morse and K. U. Ingard, *Theoretical Acoustics*. Princeton University press, 1968.

[60] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[61] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*. Springer Science & Business Media, 2013, vol. 3.

[62] D. G. Childers, D. P. Skinner, and R. C. Kemerait, "The cepstrum: A guide to processing," *Proceedings of the IEEE*, vol. 65, no. 10, pp. 1428–1443, 1977.

[63] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, "An efficient MFCC extraction method in speech recognition," in *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*. IEEE, 2006, pp. 4–pp.

[64] B. Zhen, X. Wu, Z. Liu, and H. Chi, "On the importance of components of the MFCC in speech and speaker recognition," *Acta Scientiarium Naturalium-Universitatis Pekinensis*, vol. 37, no. 3, pp. 371–378, 2001.

[65] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[66] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[67] A. V. Oppenheim, R. W. Schafer, J. R. Buck et al., *Discrete-Time Signal Processing*. Prentice Hall Upper Saddle River, 1999, vol. 5.

[68] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 5, pp. 569–586, 1999.

[69] J. Holmes, "An investigation of the volume velocity waveform at the larynx during speech by means of an inverse filter," in *Proc. Speech Communication Seminar, Stockholm*, vol. 1, 1962.

[70] S. B. Michaels, H. I. Soron, W. J. Stong, and L. P, "An accurate estimate of the glottal waveshapes," *Journal of Acoustic Society America*, vol. 23, p. 843, 1961.

[71] H. I. Michaels and W. J. Stong, "Analysis-synthesis of glottal excitation," *Journal of Acoustic Society America*, vol. 38, p. 935, 1965.

[72] W. R. Gardner and B. D. Rao, "Noncausal all-pole modeling of voiced speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 1, pp. 1–10, 1997.

[73] T. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Communication*, vol. 1, no. 3, pp. 167–184, 1982.

[74] C. H. Shadle, "The acoustics of fricative consonants," *The Journal of the Acoustical Society of America*, vol. 79, no. 2, pp. 574–574, 1986.

[75] D. O'Shaughnessy, *Speech Communication: Human and Machine*. Reading, MA: Addison-Wesley, 1987.

[76] R. L. Whitehead, D. E. Metz, and B. H. Whitehead, "Vibratory patterns of the vocal folds during pulse register phonation," *The Journal of the Acoustical Society of America*, vol. 75, no. 4, pp. 1293–1297, 1984.

[77] P. Fearnhead, "MCMC for state-space models," Lancaster University, Tech. Rep., 2011.